LA-UR-05-7070

Approved for public release;
distribution is unlimited.

Title:

MAPPIING THE STRUCTURE OF SCIENCE THROUGH USAGE

Author(s):

JOHAN BOLLEN
HERBERT VAN DE SOMPEL

Submitted to:

SCIENTOMETRICS

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Mapping the structure of science through usage.

Johan Bollen[⋆†] and Herbert Van de Sompel[†]

September 8, 2005

[†] *Research Library, Los Alamos National Laboratory, Los Alamos, NM, 87545*
[⋆] Corresponding author. Tel.: +1 505 606 0030. URL: http://public.lanl.gov/jbollen,
email: {jbollen,herbertv}@lanl.gov,

Abstract

Science has traditionally been mapped on the basis of authorship and citation data. Due to publication and citation delays such data represents the structure of science as it existed in the past. We propose to map science on the basis of usage data to determine research trends as they presently occur in both local research communities and the general scientific community. Our mapping methodology consists of a principal components analysis superimposed with a k-means cluster analysis. The subject of our analysis is a large set of usage data collected for the Los Alamos National Laboratory research community. Results indicate that meaningful maps of the interests of a local scientific community can be derived from usage data. Subject groupings in the mappings corresponds to Thomson's ISI subject categories.

Contents

1 Introduction

Science is an essential component of our globalized, technological civilization. From a social and political perspective it is therefore vital to build an understanding of its properties as a dynamic, social process. The scientific literature is rife with structural analysis and visualizations of the structure of the scientific process (Chen & Paul, 2001; visual:nagpaul2002, 2002; Boyack, 2004). The data sets used in such mappings rely mostly on citation (Leydesdorff, 2004a, 2004b), co-citation (Small, 1973; McCain, 1991; Braam, Moed, & Raan, 1991a, 1991b) and co-authorship (Wagner & Leydesdorff, 2003; He & Spink, 2002; Liu et al., 2004; Liu, Bollen, Nelson, & Sompel, 2005) data which serve as proxies to the underlying social phenomena (Everett & Pecotich, 1991). However, this focus on citation and authorship data has a number of limitations related to the nature of the publishing process:

Publication and citation delay Most peer-reviewed articles are published and cited well after they are written and submitted (Luwel & Moed, 1998; Rinia, Leeuwen, Bruins, Vuren, & Raan, 2001). Citation and authorship data will therefore reflect past scientific trends which can, only to a certain degree, be predicted from early citation data.(Adams, 2005). This is a particular problem in fast changing domains such as genomics and biochemistry.

Citation bias Citation and co-authorship are public phenomena and therefore subject to strong social desirability biases (Nederhof, 1985; King & Bruner, 2000), for example the perceived need to cite popular articles and co-author with prestigious team leaders. Such biases may obfuscate important, but implicit trends in the scientific community.

Granularity Publication is only one among many components of the scientific process. It is the end result of an iterative process which involves discussing a subject with colleagues, performing experimental research, exchanging recommendations, reading the relevant literature, investigating possible funding sources, etc. Examining the structure of science solely on the basis of publication data will provide only a highly partial picture.

There exist, however, a number of interesting precursors and proxies to publication data that merit attention in the study of scientific trends. It is commonly known that reading, publication and citation are related in the scientific process, and in particular that reading precedes publication and citation in the practices of individual scholars (Brody & Harnad, 2005). What's being read today can therefore in the aggregate serve as an indication of what will be cited tomorrow.

However, readership data is difficult to obtain and validate; how can we know whether someone has indeed read a paper? For that reason most investigations of reader-related scholarly phenomena focus their efforts on the analysis of the more general class of *usage* data. The term usage refers to a general class of information consumption and interest indicators recorded in the framework of digital information services which includes but is not limited to downloading the full-text version of a document, requesting bibliographic data, accessing a service pertaining to a particular document, etc. Each instance of usage data can with a varying degree of reliability be interpreted as an indication of user interest, and thus as an indicator of future scholarly work on the corresponding subject.

Fig. 1 shows on overview of which aspects of the scientific process usage data can capture (Egghe & Rousseau, 2000; Wouters, 1997). Since usage data can be recorded in real-time, it allows us to study the scientific process well before its results are manifested in the publication record. Furthermore, usage data is often freely available without proprietary restrictions such as those that apply to Thomson Scientific's citation data.

It can, in addition, encompass a wide range of scholarly communication items (Sompel, Payette, Erickson, Lagoze, & Warner, 2004), e.g. raw data sets and multimedia documents.

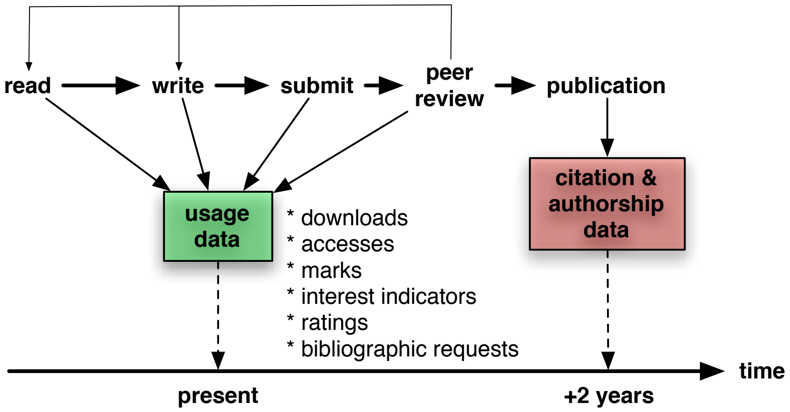


Figure 1: Usage data represents scientific trends as they occur today.

For those reasons analysis of usage data has recently emerged in efforts to detect early scientific trends. Bollen and Luce (2002) and Bollen, Sompel, Smith, and Luce (2005) discuss methods to derive metrics of journal and article impact from recorded digital library usage data. Kurtz et al. (2004a, 2004b) studies the temporal relation between reading and citation rates and formulate a model for the prediction of researcher productivity from present readership rates. Such analysis of usage data points to scientific trends as they occur in the present but are limited to one-dimensional rankings; they do not produce the structural maps of science as we commonly find in bibliometrics. As an example of the latter, Boyack (2004) produce a geographical mapping of journals using the VxInsight knowledge visualization tool (Boyack, Wylie, & Davidson, 2002) and use k-means clustering to generate subject groupings. Such visual mappings are compelling instruments to study the structure of science, but are entirely lacking in the domain of bibliometric usage studies.

Following a similar methodology as Boyack et al. (2002), we propose to create maps of science on the basis of usage data. A directed, weighted network of journal relationships was constructed from user access data recorded at the Los Alamos National Laboratory (LANL). Journal relationships are mapped in a 2-dimensional plane according to the results of a Principal Component Analysis (PCA). A k-means clustering is used to determine journal subject groupings which are overlayed with the PCA map to visualize the relations of subject domains in LANL usage patterns.

2 Methodology

Our objective is to demonstrate that usage data can be validly used to map and visualize the structure of science, much like the mappings that have previously been performed on the basis of citation data. We do so by the following three phase methodology:

Creation of usage journal networks Directed, weighted journal networks are created from usage data recorded at the LANL research library (section 2.1)

PCA mapping of journal relations A PCA analysis reduces the dimensionality of journal relationships to a set of 2-dimensional map positions. Results are validated by a visual inspection of the resulting journal positioning, an analysis of the generated principal components and the degree to which the subsequent k-means clustering overlaps with the produced visual grouping of journals (section 3).

Subject domain grouping A k-means cluster analysis groups journals according to their subject domains and is overlayed with the generated PCA map (section 3.2). The generated clusters are validated by a χ^2 analysis of how well they match Thomson's ISI subject categories (3.3).

An overview of this procedure is shown in Fig. 2. We further explain the details of this methodology below.

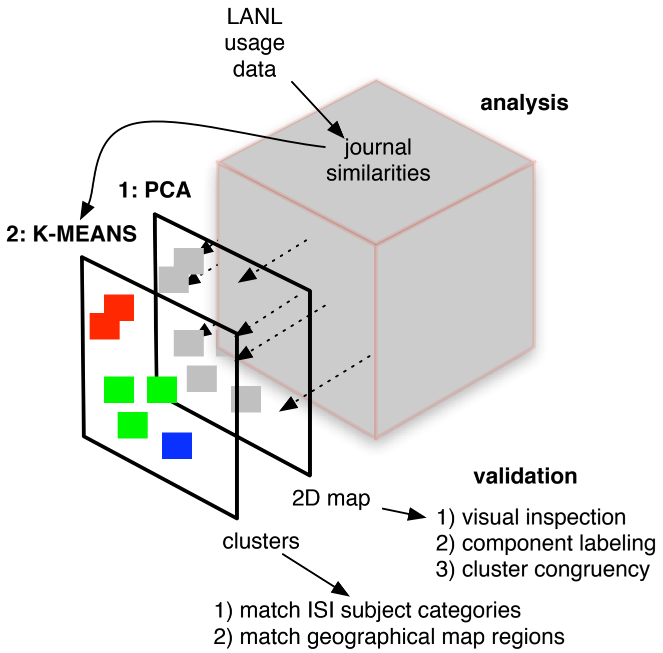
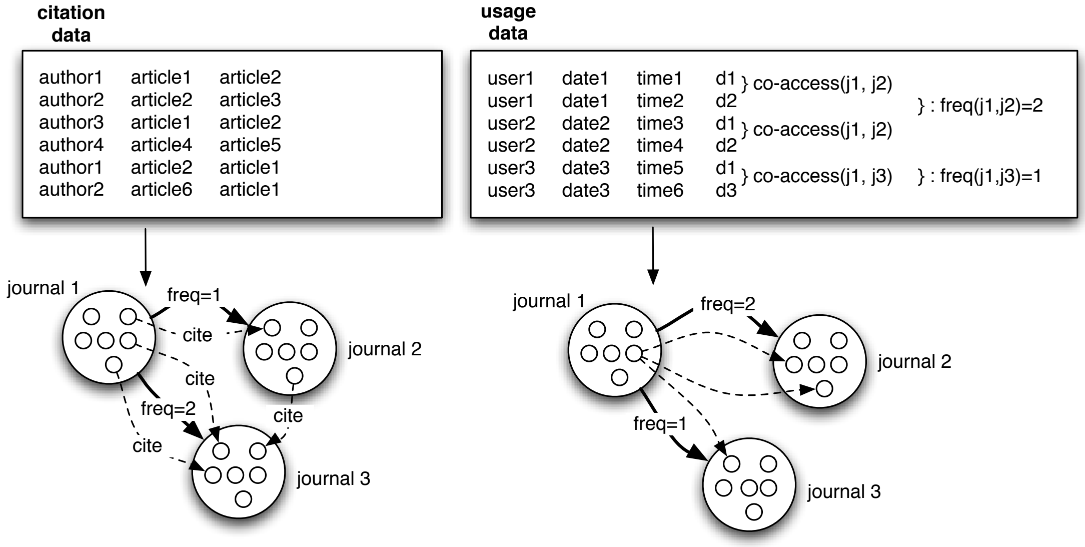


Figure 2: Mapping the structure of usage and citation by Principal Component Analysis, overlayed with k-means cluster coloring

2.1 Creation of usage networks

We extract a network of journal relationships from usage data recorded at the LANL Research Library (RL) in the period Februari 2004 to April 2005. The particular data set we used contained 392,455 document accesses, pertaining to a community of 5,866 library users, 330,109 articles and 10,696 journals. Recorded usage included a wide range of user requests such as bibliographic information requests and full-text downloads. This data set will be referred to as LANL04 for brevity.

We extracted journal relationships from the LANL04 data by a methodology outlined in (Bollen et al., 2005). The central assumption of this methodology is when users frequently access one journal article after the other this indicates the degree to which the articles, and consequently their journals, are related. The analogy to citation data is that rather than to assume two journals are related because their articles cite each other frequently, we assume that two journals are related if their articles are frequently co-accessed. Fig. 2.1 provides a graphical overview of this process which is strongly related to web usage data mining (Brin, Motwani, & Silverstein, 1997; Spiliopoulou, 1999; Chan, 1999; Srivastava, Cooley, Deshpande, & Tan, 2000; Mobasher, Dai, Luo, & Nakagawa, 2001). Although the results outlined in this paper will provide further support for the validity of the journal graphs resulting from this methodology, we refer to the cross-validation undertaken in (Bollen et al., 2005).



citation data

author1	article1	article2
author2	article2	article3
author3	article1	article2
author4	article4	article5
author1	article2	article1
author2	article6	article1

usage data

user1	date1	time1	d1	} co-access(j1, j2)	} : freq(j1,j2)=2
user1	date1	time2	d2		
user2	date2	time3	d1	} co-access(j1, j2)	
user2	date2	time4	d2		
user3	date3	time5	d1	} co-access(j1, j3)	} : freq(j1,j3)=1
user3	date3	time6	d3		

Figure 3: Extracting journal relationship data from digital library access logs

A network of journal relationships was thus extracted from the February 2004 to April 2005 LANL RL logs which as mentioned pertained to 392,455 articles which appeared in 10,696 journals. The resulting journal graph was represented by the matrix R which contained 33,256 non-zero entries. The density of the reader-based journal relationship graph was thus very low: only 3 out of 10,000 possible edges had non-zero weights.

2.2 Mapping of journal relationships

PCA and Multi-Dimensional Scaling (MDS) are well-established in scientometric research (Everett & Pecotich, 1991) and have a rich tradition in cognitive science and the social sciences (Jolliffe, 2002). PCA can extract the factors or components that best explain the variation between the items in a collection. It thus allows highly dimensional data sets to be reduced to a two or three dimensional representation based on its strongest components thereby rendering the dimensionality-reduced data set amenable to 2D or 3D plotting. In addition, the factors can be interpreted in terms of how they organize the original data set.

2.2.1 PCA mapping

It is our objective to map each journal in the LANL04 data set to positions in a 2-dimensional plane, so that similar journals are spatially close. The similarity of each pair of journals v_i and v_j was defined as the Spearman rank-order correlation coefficient ρ of their row vectors in matrix R. The Spearman rank-order correlation coefficient was chosen as a more robust, non-parametric alternative to Pearson's r in light of the derivation of matrix R from usage data. A correlation matrix R_c was subsequently created so that each of its entries $r_c(i,j) \in [-1,1]$ corresponded to the similarity of journals i and j as given by their $\rho(i,j)$. This similarity principle is similar to the notions of co-citation and bibliographic coupling (Small, 1973; Kessler, 1963).

A PCA was then performed on both matrices C_c and R_c by performing an eigenvector analysis. The two most significant components, i.e. eigenvectors with highest eigenvalues, were retained.The original journal row vectors were then projected upon these components to reduce the original $n \times n$ journal similarity matrices to a $n \times 2$ matrix denoted R'_c thus mapping each journal v_i to set of (x, y) coordinates. As such each journal could be mapped to a particular position in the Euclidean plane.

2.2.2 PCA validation

The resulting PCA map was then validated as follows:

Visual inspection We visually investigate whether the 2D positioning of journals is meaningful

Journal PCA density contour plot Journal density values in the PCA plot were visualized by means of a density contour plot to determine in which regions clustered most densely

Factor labeling The journals scoring highly on either one of the 2 principal components are examined to provide a possible semantic interpretation of the PCA factors

K-means cluster overlap Are journals within the same k-means generated clusters positioned in each other's proximity?

2.3 K-means clustering

K-means cluster analysis (Spath, 1980) consists of a greedy-algorithm which adaptively assigns items to a requested number of clusters to optimize inter-cluster distance and intra-cluster cohesion. It therefore belongs to a class of unsupervized clustering algorithms which includes Kohonen self-organizing maps (Kohonen, 1995) and automated probabilistic classifiers such as decision-tree learners (Tufekci, 2003).

2.3.1 K-means cluster method

A k-means cluster analysis was performed on the matrix R_c resulting in the assignment of each journal in the LANL04 usage networks to a particular cluster. Each journal was automatically assigned a cluster code character according to the cluster it had been assigned to, e.g. cluster 1 → "x". These codes were overlayed on top of the 2D PCA mapping of journals so that each journal had both an (x, y) coordinate in the plane and a cluster code corresponding to its cluster assignment. To demarcate the regions occupied by a particular cluster a convex hull was defined to follow the outer edges of each cluster set (same color code) in the PCA defined plane. The convex hull was rounded by a cubic spline interpolation function applied to the convex hull data points.

2.3.2 Cluster interpretation and validation

We validate the generated k-means cluster analysis on the basis of how well they match Thomson's ISI journal subject categories as follows:

Category TFIDF weighting Determine the degree to which a particular Thomson's ISI journal subject categories uniquely occurs within a particular cluster

Category distribution entropy How sharply focused or diffuse are category weights within a cluster?

χ^2 **analysis** How well do cluster assignments correspond to Thomson's ISI journal subject categories?

To interprete and validate the subject domains corresponding to a particular cluster, we extracted Thomson's ISI journal subject categories for all journals assigned to a particular cluster following (Boyack, Klavans, & Boerner, 2005). The distribution of journal subject categories could then point to an interpretation of the cluster's subject domain [1].

Not all Thomson's ISI subject categories are equally relevant to the interpretation of a particular cluster's subject domain. Some are simply more generally frequent than others. Following the principle of TFIDF index term weighting in Information Retrieval (IR) (Salton, 1998; Baeza-Yates & Ribeiro-Neto, 1999) we define a category i's weight for a given cluster j, denoted $w(i,j)$, as the ratio of the category's within-cluster frequency and its overall frequency in the LANL04 data set:

$$w(i,j) = \frac{f(i,j)}{n} \times \log\left(\frac{N_a}{n_c(i)}\right)$$

where $f(i,j)$ represents the category's raw within-cluster frequency, n the total number of categories within a cluster, N_a represents the total number of clusters, and $n_c(i)$ the category's between-cluster frequency. The resulting $w(i,j)$ values are thus the product of two factors; a category's normalized within-cluster frequency $\frac{f_c(i,j)}{f_a(i)}$ vs. its normalized between-cluster frequency, $\log\left(\frac{N_a}{n_c(i)}\right)$. This weighting scheme reduces the importance of frequent, non-cluster specific categories and increases the significance of otherwise rare, cluster-specific categories.

The "sharpness" of the weight distribution of a cluster's categories can provide information on its degree of domain focus. Following the seminal definition of entropy (Shannon, 1948) we define the category weight distribution entropy H_c as follows:

$$H_c = -\sum_i w_i \log_2(w_i) \qquad (1)$$

The entropy of a cluster's category distribution can be interpreted in a similar manner to the traditional definition of entropy in information theory: if all categories are equally strongly weighted, a cluster's subject domain is diffuse and its entropy high. Vice versa, if category weights are highly unequal the cluster's domain focus is high and its entropy it low.

Finally, we used a χ^2 analysis to determine the degree to which Thomson's ISI subject categories correspond to the generate k-means cluster assignments.

3 Results

Fig. 5 shows the results of the above described PCA and k-means analysis for the LANL04 usage data.

[1] Although it is only one among many systems of subject classification and a particularly coarse one, the ISI category system has a number of advantages: categories are manually vetted and constitute a commonly applied standard

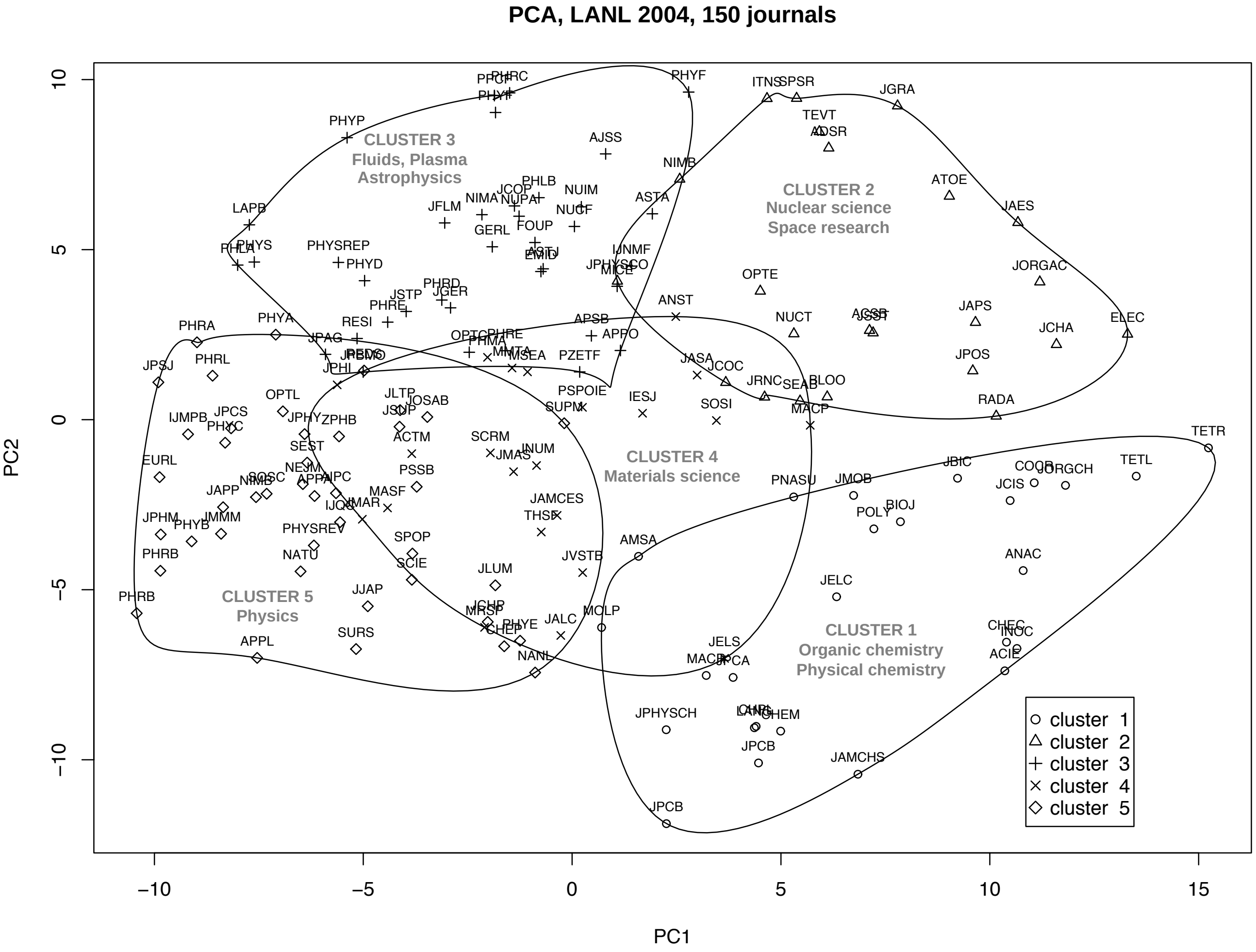


Figure 4: PCA and k-means mapping 2004 LANL journal access data.

Journal names are abbreviated to a four letter code to reduce clutter in the graph. To further increase readability the map includes only the 150 most used journals. Each cluster has been labeled with the subject domains derived in section 3.2. The PCA map and contour plot reveals a highly meaningful organization of journals, with a focus on the subject domains considered characteristic for the LANL research community. We find groupings of journals relating to physical chemistry, organic chemistry, material sciences, applied physics, plasma physics, space research and nuclear science. The structure of the generated PCA map will be discussed in more detail in the following sections.

3.1 Principal components and validation

The distributions of factor loadings for the LANL usage data reveals the degree to which LANL usage data varies according to a complex set of local constraints. Graph 6 shows the respective screeplots. The factor loadings for the 2004 LANL usage data are highly distributed, i.e. PC1 has a loading of 25% and PC2 of 17%.

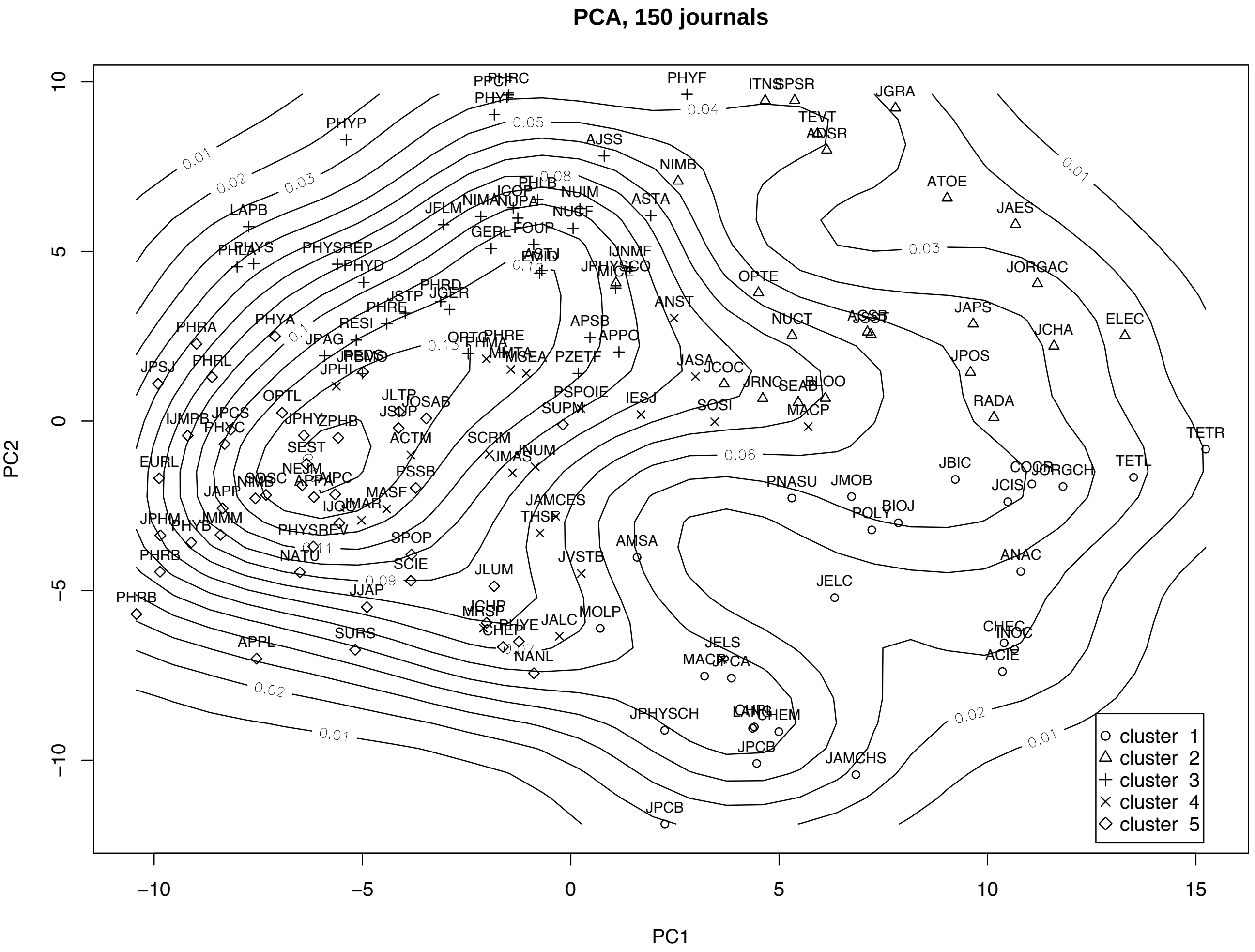


Figure 5: Contour map of spatial journal placement density in LANL04 PCA analysis.

Combined they explain only 42% of the total amount of variance among journal correlations.

We further examined the sets of journals found on the extremes of PC1 and PC2 for the results of the PCA of the 2004 LANL access data. These sets are listed in Table 1.

On the negative end of PC1 we find a preponderance of journals relating to physics, in particular those relating to condensed matter. On the positive extreme, we find a majority of journals relating to chemistry, in particular organic chemistry. This corresponds to the content of the clusters that are located on the extremes of PC1, shown in Fig. 5, namely on the positive extreme cluster 1 and 2 (organic chemistry and biology) and on the negative extreme of PC1 cluster 5 (physics). This suggests an interpretation in terms of a split between natural and life sciences, except that in this case the split seems to be related to "inorganics" vs. "organics", indicative of research community with a more narrow focus on the natural sciences, in particular physics and chemistry.

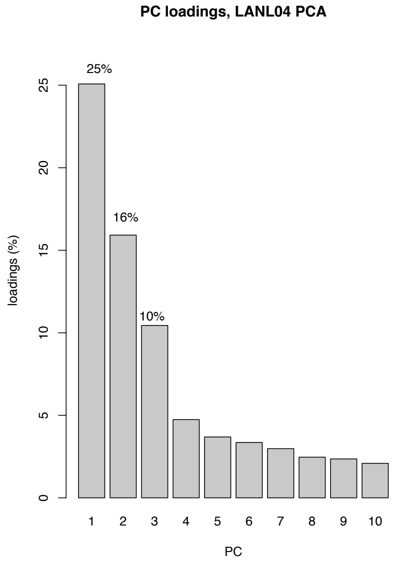


Figure 6: Factor loadings of PCA performed for 2004 LANL usage data

The second component PC2 seems to represent a more elusive dimension of our usage data. The journals on the negative extreme of PC2 correspond to chemistry but with a particular focus on chemical physics, colloids and surfaces. The positive extreme of PC2 corresponds to plasma physics, space research including geophysics and nuclear physics. Along the axis of PC2 we see a gradual shift from physical chemistry to material science to plasma, fusion, fluids and astrophysics which seems to suggest a transition from the chemical "micro" constituents of matter, note the journal "Nano Letters", to studies of "macro" phenomena relating to geophysics, space research, and astrophysics. Unfortunately, an adequate semantic label for this shift eludes us.

Of particular interest is the position of cluster 4 (materials science) at the intersection of PC1 and PC2. Material science can indeed be situated on the intersection of chemistry and physics, and the "micro" world of physical chemistry to the "macro" domain of fusion, plasma research, and astrophysics.

3.2 Usage clusters: category weighting and entropy

The k-means clustering of LANL usage (Fig. 5) and JCR03 citation data yielded 5 distinct clusters. The information entropy was calculated for each cluster's category distributions to determine the degree to which a particular cluster corresponded to a particular well-defined domain and compare the domain focus usage and citation clusters.

Most category distributions are characterized by a sharp decline of category weight values after the first 2 or three categories and thus correspond narrowly to a particular subject domain as shown in Fig. 9 (appendix). Using the 3 or 4 most highly valued journal categories, we subjectively label the generated cluster content as shown in Table 2. Category weight distribution entropy values are listed after the cluster labels as an indication of the cluster domain focus.

We find a patterns that separate physics and related domains from chemistry and biology. As indicated by the contour map of journal density at the bottom of Fig. 5 the LANL04 clusters 3, 4, and 5 form a conglomerate

LANL04 - PC1 (horizontal)		LANL04 - PC2 (vertical)	
PC1 < -8	PC1 > 10.4	PC2 < -7.2	PC2 > 7.7
PHYS REV B	INORG CHEM	CHEM PHYS LETT	PHYS PLASMAS
PHYS REV LETT	TETRAHEDRON LETT	J PHYS CHEM B	PHYS FLUIDS
PHYS REV A	TETRAHEDRON	J PHYS CHEM A	IEEE T NUCL SCI
PHYSICA B	J ORGANOMET CHEM	CHEM MATER	ADV SPACE RES
J PHYS-CONDENS MAT	ELECTROPHORESIS	J AM CHEM SOC	PHYS FLUIDS
J APPL PHYS	ANAL CHEM	LANGMUIR	J GEOPHYS RES ATMOS
PHYSICA C	J ORG CHEM	J PHYS CHEM	SPACE SCI REV
J MAGN MAGN MATER	J AEROSOL SCI	ANGEW CHEM INT EDIT	PLASMA PHYS CONTR F
EUROPHYS LETT	COORDIN CHEM REV	J PHYS CHEM B	PHYS REV C
PHYS LETT A	J COLLOID INTERF SCI	MACROMOLECULES	ASTROPHYS J SUPPL S

Table 1: Journals on negative and positive extremes of PC1 and PC2 resulting from PCA for 2004 LANL access data.

LANL 04

Cluster	Label	Category codes	Entropy
1	**Organic Chemistry and biology**:	EI, EE, DY, CQ, UY	$H_c = 3.411$
2	**Nuclear Science**:	RY, EA, EC, CO	$H_c = 4.365$
3	**Fluids, plasma and astrophysics**:	UF, UI, UR, UN, PU, SY, BU	$H_c = 4.029$
4	**Materials science**:	PM, PZ, UP, UK	$H_c = 3.261$
5	**Condensed matter and physics**:	UK, UB, UH, UI, SY	$H_c = 3.517$

Table 2: Labels assigned to LANL04 clusters on the basis of most highly weighted cluster ISI subject categories

of natural science research focused on materials science, applied physics, condensed matter and fluids and plasma. On the right of the LANL04 PCA map, Fig. 5, we find a less cohesive aggregation of the journals in clusters 1 and 2 which relate to nuclear science, inorganic chemistry and organic chemistry and biology. Of particular interest are a group of journals in cluster 2 relating to ongoing research on nuclear propulsion and applications in space. In addition, in cluster 3 we find a group of journals on the subject of applications of nuclear physics to astronomy and geophysics indicating the existence of a research group involved with astrophysical models for space observation and modeling. Cluster 4, material science, overlaps with all other clusters thereby indicating its multi-disciplinary focus at the intersection of physics, chemistry, and to a lesser degree organic chemistry and nuclear science.

3.3 Cluster validation

To validate the generated k-means clusters we performed a χ^2 analysis (Sheskin, 2004) which tested whether k-means cluster assignments and journal ISI categories were significantly related, i.e. did usage clusters overlap with journals' ISI categories? In case the χ^2 analysis indicated a significant relationship, we calculated Cramer's Phi coefficient (Cramer, 1946), denoted ϕ_C to assess the strength of that relationship [2]. Table 3 lists the results of this analysis.

[2] Cramer's ϕ_C varies in the $[0, +1]$ interval where 0 indicates the absence of a relationship, +1 indicating a strong relationship

LANL 04

χ^2	df	p
293.610	140	< 0.001
$\phi_C = 0.699$		

Table 3: Chi-square analysis for cluster assignment and ISI categories in JCR03 and LANL04 data sets

In both cases the relationship between k-means cluster assignments and the journal's ISI categories are highly statistically significant, i.e. $p < 0.001$, meaning that the LANL04 usage clustering corresponds well to the ISI categorization. In addition, the calculated ϕ_C values were found to be 0.699 indicating a strong relationship between k-means clustering and ISI categories.

4 How about citation?

We applied the discussed methodology to the 2003 Thomson's ISI Journal Citation Reports (Science Edition) to validate its ability to map the structure of scientific domains. In this case, however, the resulting mapping would reflect the structure of science for the community of authors publishing in the set of Thomson ISI selected journals.

A journal relationship matrix was derived from the 2003 Thomson's Journal Citation Reports for all pairs of journals in the collection (8624). For each pair of journals A and B we obtain a citation count which corresponds to the frequency with which articles in journal A published in 2003 cited articles published in journal B for the two preceding years (2002 and 2001). A 8624 journal graph resulted which we represented by the matrix C which contained 1,004,289 non-zero entries, indicating a highly sparse journal citation graph: only 1% of all possible entries of matrix C had non-zero weights.

The resulting PCA mapping is shown in Fig. 8. Table 4 lists the cluster label generated on the basis of the discussed category frequency weighting procedure and the corresponding entropy values. In addition, the χ^2 analysis results for the comparison of JCR03 k-means cluster assignment and Thomson's ISI journal domain subject categories are listed in Table 5.

JCR03

Cluster	Label	Category Codes	Entropy
1	**Astronomy and astrophysics**:	BU, UP, UN, UI, SY	$H_c = 2.524$
2	**Biochemistry and cell biology**:	CQ, DM, RU, NI, CU	$H_c = 4.276$
3	**Physics and geophysics**:	UK, UB, UI, UH, LE, FI	$H_c = 3.218$
4	**Medicine and microbiology**:	MA, DE, WE, YQ, YP, QU	$H_c = 4.557$
5	**Chemistry and materials science**:	EE, DY, EC, EI, EA, UY	$H_c = 3.232$

Table 4: Labels assigned to clusters on the basis of most highly weighted cluster ISI subject categories for JCR03 data.

We observe a more dense, clustered placement of journals in the JCR03 PCA mapping as well as the contour plot. The particular journal clusters listed in Table 4 correspond to more general, less institution-driven

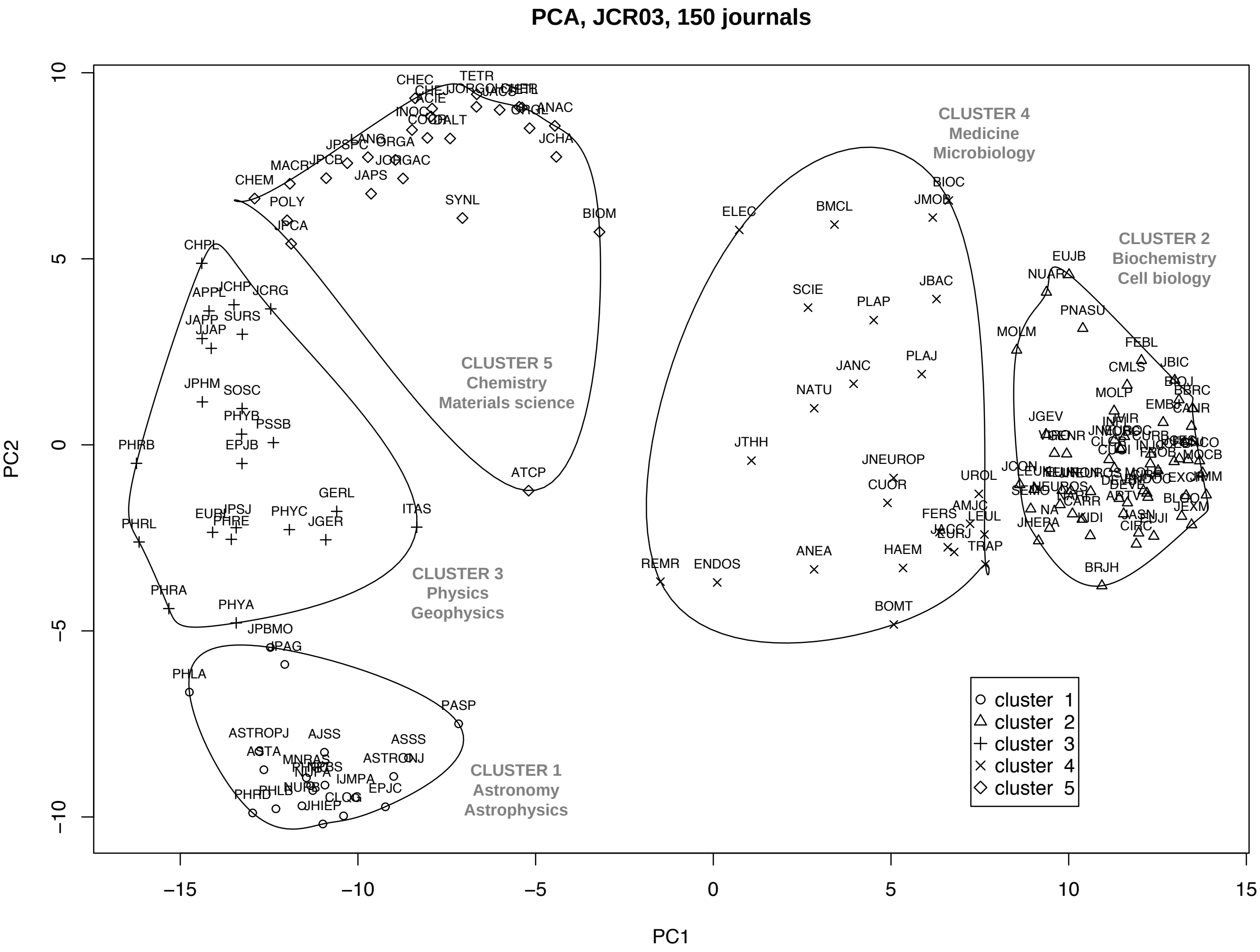


Figure 7: PCA and k-means cluster model of 2003 Journal Citation Reports data, including a contour map of spatial journal placement density.

JCR03

χ^2	df	p
469.000	172	< 0.001
$\phi_C = 0.884$		

Table 5: Chi-square analysis for cluster assignment and ISI categories in JCR03 data set.

subject categories such as "astronomy and astrophysics", whereas the LANL04 clustering is strongly driven by institutional focii such as "applications of plasma physics to astrophysics". In spite of the denser arrangement of journals, the JCR03 clusters are not more strongly focused on a particular subject domain. Entropy values for the JCR03 clusters do no significantly differ from entropy values for the LANL04 cluster as evidenced by

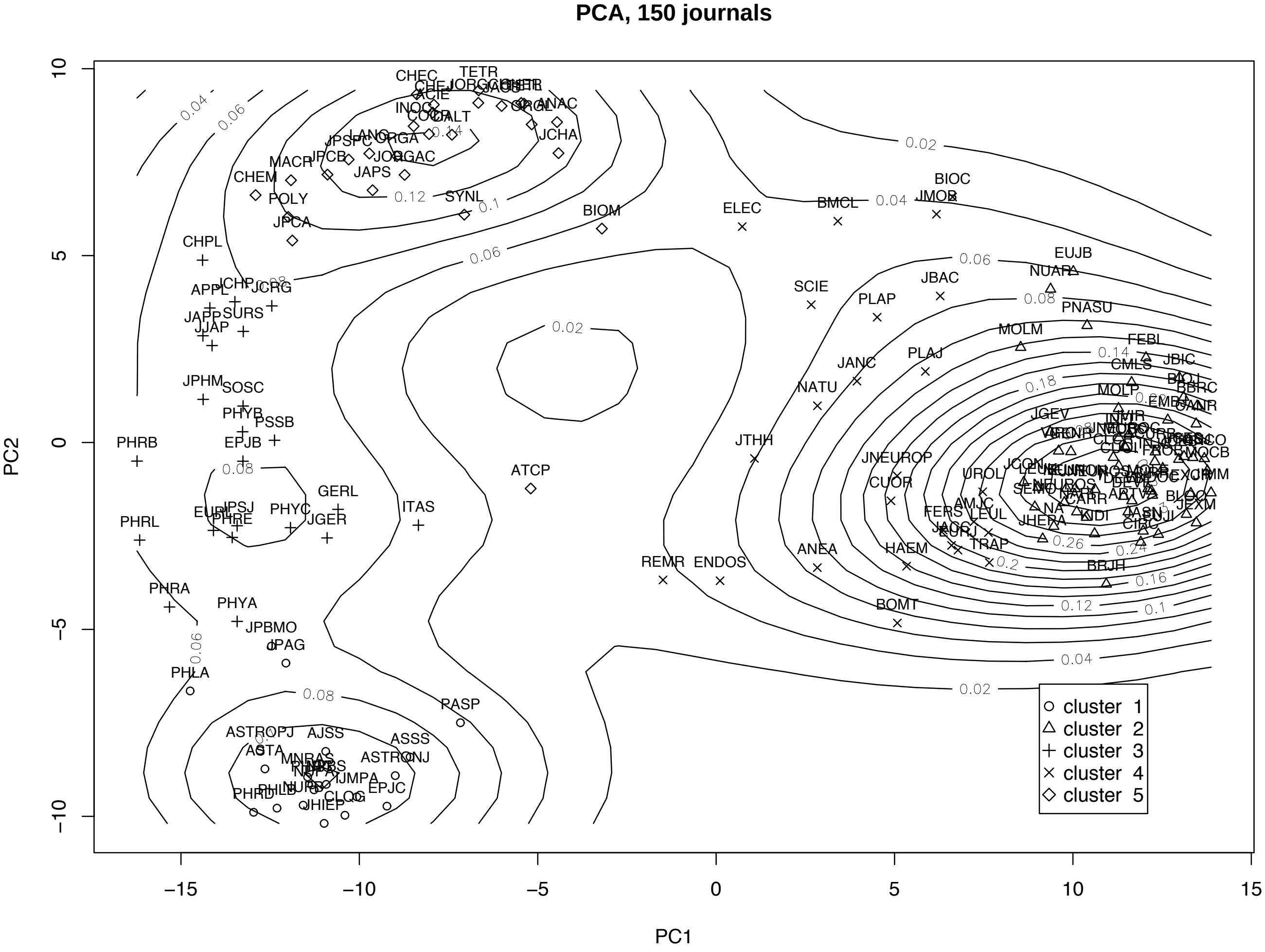


Figure 8: PCA and k-means cluster model of 2003 Journal Citation Reports data, including a contour map of spatial journal placement density.

a t-test ($df = 5, p = 0.73$).

As is the case for the LANL04 data, the JCR03 k-means clustering is statistically significantly related to Thomson's ISI subject categories. However, the JCR03 ϕ_C is slightly higher, i.e. 0.884 vs. 0.699, indicating citation patterns correspond better to existing subject categories.

These results indicate that the applied methodology can identify valid journal maps and subject groupings on the basis of citation data, thereby validating its use on the LANL04 usage data and the subsequent analysis results.

5 Conclusion

Although there exist many methodologies to map the structure of science, most are reliant on the use of citation and authorship data, i.e. they use the structure of the body of published material. Due to publication delays and citation biases any investigation of the structure of science on the basis of citation data is essentially studying science at is was 2 to 3 years ago. Since present usage rates have been shown to predict future citation rates, it has been speculated that usage data can be used as a viable, more contemporary proxy to scientific trends.

We have formulated a methodology to reconstruct networks of document relationship from access patterns recorded in DL log data. We have done so for a set of LANL RL logs recorded for one year through 2004 and 2005. A network of 10,000 journals was constructed and subjected to a PCA overlayed with a 5 cluster k-means analysis.

We conclude the following from these results. First, the combined PCA and k-means clustering visualization indicate that a meaningful organization of journal clusters and components can be reconstructed on the basis of usage data. Not only can usage data be applied to the ranking of journals and articles on the basis of usage or access frequency, the temporal patterns by which they are accessed contain information from which reliable and meaningful relational data can be inferred.

Second, when we examine the differences between the usage and citation PCA mapping and k-means clustering we observe that:

1. The geographical distribution of journals according to the PCA performed on usage data is more diffuse than the distribution of journals according to the PCA performed on citation data.
2. Journal title terms in the generated usage k-means clusters are equally strongly focused on a set of particular subject domains relevant to the LANL research community
3. Usage and citation cluster do not overlap in terms of subject domains. Usage defines a grouping of journals and subjects which is shaped by local institutional foci and contingencies.

Citing behavior may be subject to a smaller number of stronger factors, i.e. citers are guided by fewer but stronger semantic dimensions such as "life" vs. "natural sciences. Indeed, a comparison of the distribution of factor loadings reveals that LANL usage is shaped by a more diffuse set of factors than the JCR citation data. Journal correlations in the 2003 JCR PCA can be largely explained by the first two components (87%) while the LANL PCA's first 2 factors explain only 42% of all journal correlations. Citers furthermore adhere to commonly accepted subject groupings, such as physics, chemistry, astronomy and biology. Usage seems to be more diffuse and interdisciplinary, and less easily explained by existing subject categories. As an example of such interdisciplinary focus we refer to the observed LANL04 journal clusters related to applications of nuclear energy to space technology, and a materials science cluster which overlaps with a range of clusters on subjects as varied as organic chemistry, plasma physics and condensed matter.

This paper has presented a highly preliminary attempt to compare the components and structure of usage and citation behavior in the scientific community. In terms of our data samples and methodology much work remains to be done. Although the 2004 and 2005 LANL usage data provided an extensive sample of journal usage patterns in the LANL research community, it does not represent usage in the scientific community. Therefore our results confound both LANL specific patterns and general characteristics of reading behavior. To fully represent the scientific community and the characteristics of reading behavior as compared to citation behavior we need to expand our log data sample to a wide range of representative institutions. Although certain technical hurdles exists in the large-scale aggregation of log data, Sompel, Young, and Hickey (2003) discusses a number of practical options.

The limitation of sample size and coverage applies equally to our citation data. We used Thomson's 2003 JCR data, but this applies to a small subset of all published literature, i.e. a selection of journals for which Thomson has decided to publish citation and impact data. Possible future extensions of this work may rely on public sources of citation data such as Citeseer or arXiv to increase the relevance and scope of this work as shown by Boyack (2004). In addition, the discussed analysis should be extended to the level of individual articles for a finer grained map of science. Citation and usage data are widely available on the article level and the analysis described in this paper can thus be applied to such finer-grained, article level data.

Since usage data is generally more current than citation data, an interesting future research area would be to study the evolution of usage clusters and components over time. Given sufficient longitudenal log data reflecting the usage patterns of a representative sample of the scientific community, it is entirely feasible to study the temporal evolution of interest clusters and use such data to predict future scientific trends. The factor loading of our PCA analysis suggests furthermore that 3D models of usage and citation could be constructed by including additional components.

We conclude from these results that usage data is a viable and essential part of the future scientometric instrumentarium. The discussed mapping of science on the basis of DL log data offers tantalizing clues to the possibility of studying local and global scientific trends as they occur in the present, free from the biases, delays and proprietary decision-making that generally accompanies citation and authorship data.

Acknowledgements

We want to thank Marko Rodriguez for his relentless proof-reading and contributions to the paper. Furthermore, we wish to extend our gratitude to the R developers (`http://www.r-project.org`) whose statistical software allowed us to perform the PCA and clustering analysis.

References

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, *63*(3), 567–581.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern information retrieval.* ACM Press / Addison-Wesley.

Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, *8*(6).

Bollen, J., Sompel, H. V. de, Smith, J., & Luce, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, *In press*.

Boyack, K. W. (2004). Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1).

Boyack, K. W., Klavans, R., & Boerner, K. (2005). Mapping the backbone of science. *Scientometrics*, *In press*.

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using vxinsight(r) for science and technology management. *Journal of the American Society for Information Science and Technology*, *53*(9), 764–774.

Braam, R. R., Moed, H. F., & Raan, A. F. J. van. (1991a). Mapping of science by combined co-citation and word analysis. I. structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251.

Braam, R. R., Moed, H. F., & Raan, A. F. J. van. (1991b). Mapping of science by combined co-citation and word analysis. II: dynamical aspects. *Journal of the American Society for Information Science*, *42*(4), 252-266.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the 1997 acm sigmod international conference on management of data* (pp. 265–276). ACM Press.

Brody, T., & Harnad, S. (2005). *Earlier web usage statistics as predictors of later citation impact* (Eprint No. 10647). ECS, Intelligence, Agents, and Multimedia Group: University of Southampton.

Chan, P. K. (1999). Constructing web user profiles: a non-invasive learning approach. In B. Masand & M. Spiliopoulou (Eds.), *Web usage analysis and user profiling - LNAI 1836.* San Diego, CA: Springer.

Chen, C. M., & Paul, R. J. (2001). Visualizing a knowledge domains intellectual structure. *Computer*, *34*(3).

Cramer, H. (1946). *Mathematical models of statistics.* Princeton, NJ: Princeton University Press.

Egghe, L., & Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for information science*, *51*(2), 158–165.

Everett, J. E., & Pecotich, A. (1991). A combined loglinear/MDS model for mapping journals by citation analysis. *Journal of the American Society for Information Science*, *42*(6), 405–413.

He, S., & Spink, A. (2002). A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. *Journal of the American Society for Information Science and Technology*, *53*(11), 953–959.

Jolliffe, I. T. (2002). *Principal component analysis.* New York: Springer Verlag.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*, 10-25.

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing*, *17*(2), 79–103.

Kohonen, T. (1995). *Self-organizing maps.* Berlin: Springer.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., & Murray, S. S. (2004a). The bibliometric properties of article readership information. *JASIST*, *56*(2), 111–128.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., & Murray, S. S. (2004b). Worldwide use and impact of the NASA Astrophysics Data System digital library. *JASIST*, *56*(1), 36–45.

Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in journal citation reports. *The journal of documentation*, *60*(4), 371–427.

Leydesdorff, L. (2004b). Top-down decomposition of the journal citation report of the social science citation index: graph- and factor-analytical approaches. *Scientometrics*, *60*(2), 159–180.

Liu, X., Bollen, J., Nelson, M. L., & Sompel, H. V. de. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, *In press.*

Liu, X., Bollen, J., Nelson, M. L., Sompel, H. V. de, Hussell, J., Luce, R., & Marks, L. (2004). Toolkits for visualizing co-authorship graph. In *Proceedings of the 2004 joint acm/ieee conference on digital libraries* (pp. 404–404). Tuscon, AZ.

Luwel, M., & Moed, H. F. (1998). Publication delays in the science field and their relationship to the ageing of scientific literature. *Scientometrics*, *41*(1–2), 29–40.

McCain, K. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for Information Science*, *42*(4), 290-296.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data. In *Widm '01: Proceedings of the 3rd international workshop on web information and data management* (pp. 9–15). ACM Press.

Nederhof, A. J. (1985). Methods of coping with social desirability bias - A review. *European Journal of Social Psychology*, *15*(3), 263–280.

Rinia, E. J., Leeuwen, T. N. van, Bruins, E. E. W., Vuren, H. G. van, & Raan, A. F. J. van. (2001). Citation delay in interdisciplinary knowledge exchange. *Scientometrics*, *51*(1), 293–309.

Salton, G. (1998). Term-weighting approaches in automatic text retrieval. *Information processing and management*, *24*(5), 513–523.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Sheskin, D. J. (2004). *Parametric and nonparametric statistical procedures.* New York, NY: Chapman and HallCRC.

Small, H. (1973). Co-Citation in the scientific literature: a new measure of the relationship between documents. *Journal of the American Society for Information Science*, *42*, 676–684.

Sompel, H. V. de, Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication - Building the system that scholars deserve. *D-Lib Magazine*, *10*(9).

Sompel, H. V. de, Young, J. A., & Hickey, T. B. (2003). Using the OAI-PMH ... Differently. *D-Lib Magazine*, *9*(7-8).

Spath, H. (1980). *Cluster analysis algorithms.* Chichester, UK: Ellis Horwood.

Spiliopoulou, M. (1999). The laborious way from data mining to web mining. *Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web".*

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, *1*(2), 12–23.

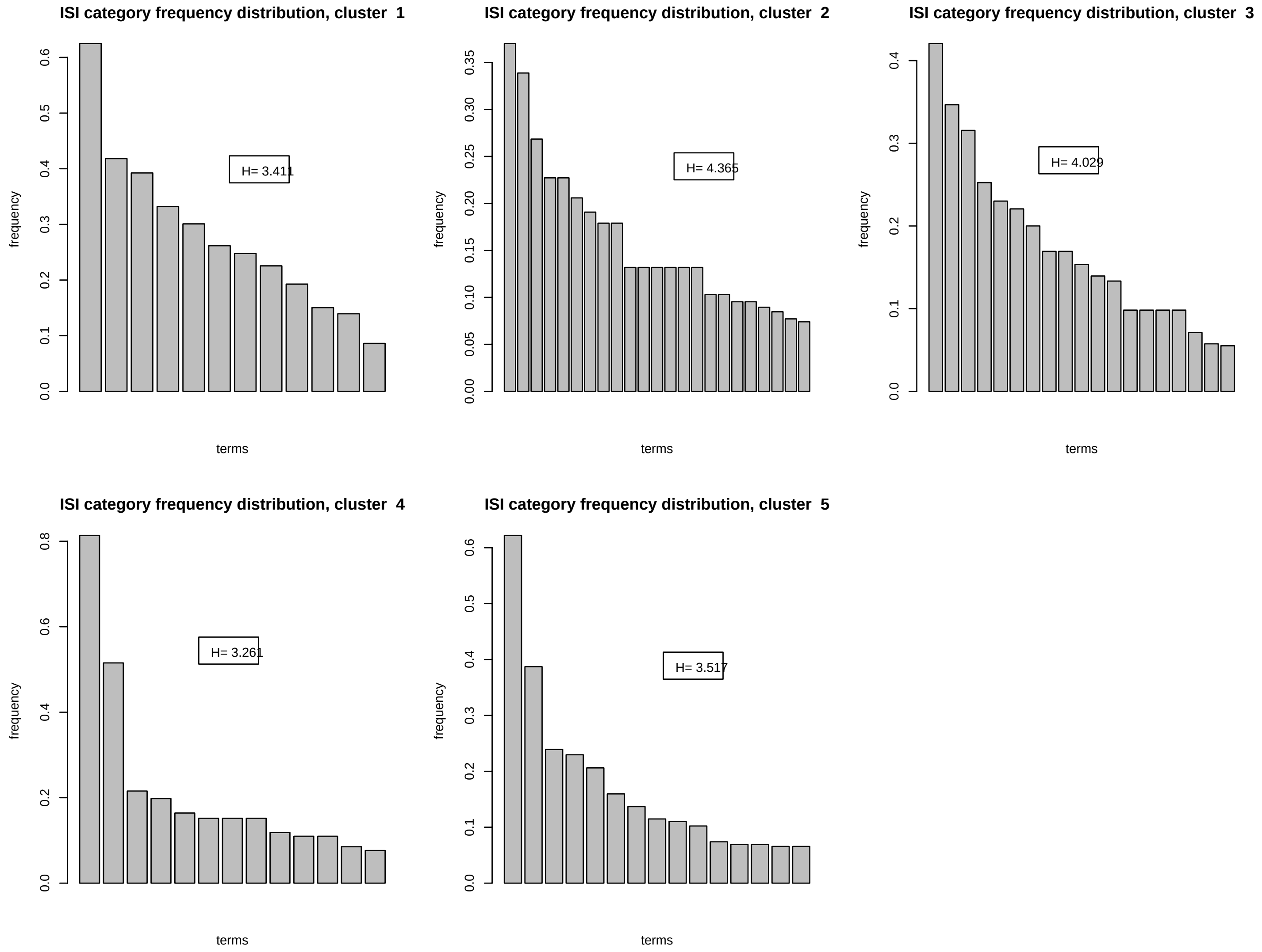
Tufekci, S. (2003). Generalized decision trees: methodology and applications. *Computers and industrial engineering*, *24*(1).

Visualizing cooperation networks of elite institutions in india. (2002). *Scientometrics*, *54*(2), 213–228.

Wagner, C. S., & Leydesdorff, L. (2003). Mapping global science using international co-authorships: a comparison of 1990 and 2000. In *Ninth international conference on scientometrics and informetrics.* ISSI.

Wouters, P. (1997). Citation cycles and peer review cycles. *Scientometrics*, *38*(1), 39–55.

Appendix



ISI category frequency distribution, cluster 1

ISI category frequency distribution, cluster 2

ISI category frequency distribution, cluster 3

ISI category frequency distribution, cluster 4

ISI category frequency distribution, cluster 5

Figure 9: Cluster specific journal classification distributions for LANL04 data.